On various ways of tackling incomplete information in statistics

Didier Dubois

A B S T R A C T

This short paper discusses the contributions made to the featured section on Low Quality Data. We further refine the distinction between the ontic and epistemic views of imprecise data in statistics. We also question the extent to which likelihood functions can be viewed as belief functions. Finally we comment on the data disambiguation effect of learning methods, relating it to data reconciliation problems.

Keywords:
Incomplete information
Random sets
Fuzzy sets
Evidence theory
Imprecise probability
Possibility theory

1. Introduction

The set of position papers gathered in the special section on low quality data proposes various ways of handling incomplete information in statistics. Imprecision may pervade the chosen model or the observed data. Moreover, prior information is generally poor. Two of these contributions focus on imprecise data, two other ones on the lack of prior information. On the one hand, one question is whether set-valued data can be handled just like any other kind of complex data. On the other hand, there is the problem of choosing a formal framework for handling incomplete information. The paper cosigned by this discussant [3] shares with the paper on fuzzy random variables [21] the use of multiple-valued mappings and random sets, but the way proposed to exploit set-valued data is radically different, as further discussed in the next section. Another pair of strikingly different papers dealing with related issues is formed by Denoeux and Masegosa–Moral papers that deal with the role of likelihood functions when prior information is poor and cannot be modelled by a unique probability distribution on the parameter space. In [14], what is proposed is essentially a form of sensitivity analysis over Bayesian inference, where the likelihood function alone is considered as totally insufficient to allow for any form of learning. However, Denoeux [4] argues to the converse, namely by exploiting an idea originally proposed by Shafer in his book [16]: he claims that the likelihood function does inform us to some extent on the value of the parameter of a model, when an observation becomes available. It is then possible to handle fuzzy data as well. Finally the paper by Huellermeier [12], even if it does not use likelihood functions explicitly (but he shows how they can be laid bare), suggests that the chosen class of models may help reducing the imprecision of the data. In the following we briefly comment these contributions.

2. On the distinction between ontic and epistemic data

In our position paper [3], we made the distinction, also endorsed by Huellermeier [12], between ontic and epistemic views of fuzzy set-valued data. The impressive set of statistical methods developed by the Oviedo SMIRE team [21] considers fuzzy set-valued data as precise entities belonging to a space of functions [11], equipped with suitable operations in order to preserve the fuzzy set semantics of such functions (especially fuzzy arithmetic operations). As a consequence, while the

http://dx.doi.org/10.1016/j.ijar.2014.04.002

mean value in this setting is a fuzzy interval, the variance is precise (based on sophisticated distances between fuzzy sets). In this sense, the view of fuzzy data advocated by this group is clearly ontic. However, the applications they developed elsewhere (such as human perception of length, and flood prediction [2]) handle low quality human-originated data on numerical quantities; to quote them [21]: "variables or attributes [that] can only be observed imprecisely". However the ontic view of set-valued data is primarily devoted to natural entities that take the form of fuzzy sets and that are tainted with variability: observations of regions in an image, time intervals during which some activities take place, blood vessel snapshots, vectors of performance ratings across a population of candidates, etc.

But things are not that simple. In fact, the ontic–epistemic distinction does not correspond to the objective–subjective distinction exactly, that is, ontic set-valued data may not just reflect sets that occur as such in the nature. There are circumstances when epistemic set-valued data, even if they are imprecise descriptions of otherwise point-valued variables can be treated as ontic entities. Here are two examples:

- Suppose one imprecisely measures a precisely defined attribute a number of times, say via a number of different observers (e.g. human testimonies on the value of a quantity), but the actual aim of the statistics is to model the variability in the imprecision of the observers. In other words, while such fuzzy data are subjective descriptions of an otherwise objective quantity, they can be considered as ontic with respect to the observers. In particular, if agreeing observers provide nested set-valued estimates of a constant but ill-known value (with various degrees of confidence), one may consider that the various levels of precision correspond to a form of variability, which justifies the use of a scalar distance between such sets in the computation of the variance. But it is the variance of the imprecision levels of the observer responses that is obtained. This variance says little about the properties of the objective quantity on which observers report.
- Sometimes human perceptions refer to a complex matter that cannot be naturally represented by a precise numerical value. For instance, ratings in a dish tasting experiment are verbal rather than numerical. Imprecise terms then refer to no clear objective feature of the phenomenon under study. For instance, the taste of a dessert does not directly describe the objective ingredients of the dish. So, the collected imprecise data can be considered ontic, because you want to know if people will like the dessert, not how much butter or sugar it contains. Here again, the human perceptions can be handled as ontic entities. However, the question is then to figure out whether the collected subjective data in this kind of situation is liable of a representation by means of a fuzzy set over a numerical scale: indeed, the very reason why a precise numerical estimate is inappropriate in this kind of situation is because a one-dimensional numerical scale does not make sense. Then, the statistician is not better off when representing human perceptions by means of fuzzy sets (let alone trapezoidal ones) on a meaningless numerical scale.

In summary, a set-valued statistic is ontic if the set representation captures the essence of the issue under study; it is epistemic if the purpose is to provide some information on the precise entity that could not be precisely observed because of the poor quality of the knowledge. Adopting an ontic approach to the statistics of human perceptions of otherwise objective quantities yields a description of the observer behaviour, not of the natural phenomenon on which this observer reports.

3. On likelihoods in the contexts of belief functions, possibility, and imprecise probability theories

In his position paper, Denoeux [4] argues in favour of the use of likelihood functions for building data-driven belief functions, assuming the contour function of the belief function should be taken as proportional to the likelihood function, thus following a suggestion made in Shafer's book [16].

This point of view is strengthened by the formal result proved in the paper, namely that this approach is the consequence of the likelihood principle, the consistency with Bayes rule in case a probabilistic prior is available, and the minimal commitment principle: there exists a unique minimally committed belief function whose contour function is proportional to the likelihood function $L(\theta) = P(x|\theta)$, where x is the observation, and $\theta \in \Theta$ is the parameter of the distribution generating x, and this belief function is consonant. In other words, it is a necessity measure based on the possibility distribution

$$\pi(\theta) = \frac{L(\theta)}{\max_{\tau \in \Theta} L(\tau)}.$$

This result looks compelling. However, it relies on an assumption that may sound questionable, namely that the relative information contained in belief functions is evaluated on the basis of their commonalities. This view has been advocated quite early by Smets [18], but there are alternative definitions of relative information to the comparison of commonalities, such as specialisation (random set inclusion) and the comparison of plausibility (or belief) functions [8,20]. Interestingly, in [20], Smets advocates the latter approach as the basis for the minimal commitment, not the comparison of commonalities. He also spends some time discussing the concept of specialisation, but seems to have given up using commonality functions. It suggests that the issue of choosing a proper information comparison technique for belief functions was not quite settled in his mind. In the following discussion, we focus on the choice between the comparison of plausibilities, and of commonalities on which Denoeux relies.

Table 1
Example of incomparable belief functions.

0	*ab*	*ac*	*bc*
Q_1	0.5	0.5	0
Q_2	0.5	0.5	0.5
Pl_1	1	1	1
Pl_2	1	1	0.5

A belief function is defined by a basic belief mass (bbm) assignment function $m : 2^\Theta \to [0, 1]$ such that $\sum_{E\subseteq\Theta} m(E) = 1$, and we assume $m(\emptyset) = 0$. Then the belief, plausibility and commonality functions are respectively

$$Bel(A) = \sum_{E\subseteq A} m(A); \qquad Pl(A) = 1 - Bel\big(A^c\big); \qquad Q(A) = \sum_{E\supseteq A} m(A).$$

A mass function m_1 is said to be

- less committed than m_2 in the wide sense if and only if for all $A \subseteq \Theta$, $Q_1(A) \geq Q_2(A)$.
- less informative than m_2 in the wide sense if and only if for all $A \subseteq \Theta$, $Pl_1(A) \geq Pl_2(A)$, or equivalently $Bel_1(A) \leq Bel_2(A)$.

The two terms "committed" and "informative" are used here for the sake of clarity and convenience. Both definitions make sense, since

- $Q(A)$ is all the greater as masses are assigned to larger subsets of Θ which makes the bbm less committed to specific elements, hence closer to the vacuous belief function.
- m_1 is less informative than m_2 means that the intervals $[Bel_1(A), Pl_1(A)]$ contain $[Bel_2(A), Pl_2(A)]$, that is, m_1 clearly gives less information on the likelihood of events than m_2. In particular the set of probability functions compatible with m_1 is larger than the one compatible with m_2.

To-date, it is not entirely clear which of the two information orderings is better adapted to the purpose of comparing belief functions by their informational content. This is troublesome as the two orderings are not consistent with each other: one may have $Q_1 > Q_2$ along with $Pl_1 < Pl_2$, and this type of phenomenon will precisely occur if the contour functions of m_1 and m_2 are equal. The simplest example is known.

Example. Consider $\Theta = \{a, b, c\}$, $m_1(\{a, b\}) = m_1(\{a, c\}) = 0.5$, $m_2(\{a\}) = m_2(\{a, b, c\}) = 0.5$. It is clear that both bbm's have the same contour function $\pi(a) = 1, \pi(b) = 0.5, \pi(c) = 0.5$. Moreover, as Table 1 shows, it is clear that m_1 is strictly more committed than m_2, but also strictly less informative than m_2.

Denoeux bases his use of belief functions and possibility distributions in statistical applications on the following mathematical result that we reformulate outside the context of likelihood functions:

Proposition. *The unique least committed belief function in the set of belief functions having a contour function $C_m(\theta) = \sum_{E:\theta\in E} m(E)$ proportional to a given (normalised) possibility distribution π is the necessity measure N induced by π (and defined by $N(A) = \min_{\theta\in A} 1 - \pi(\theta)$).*

This result is closely related to the one by Smets [10] whereby the least committed belief function among those that are more committed than two consonant belief functions is the consonant belief function induced by the pointwise minimum of the two contour functions induced by each consonant belief function (just assume the two original consonant belief functions are the same). It justifies the minimum rule of possibility theory.

However, using this result to justify the interpretation of a likelihood function as a consonant belief function may look problematic: it is well-known [9] that if a belief function defined by the bbm m has a *normalised* contour function C_m associated to possibility and necessity measures $\Pi_m(A) = \max_{\theta\in A} C_m(\theta)$, and necessity measure $N_m(A) = 1 - \Pi_m(A)$, then in general $[N_m(A), \Pi_m(A)] \subset [Bel(A), Pl(A)]$, which suggests that the consonant approximation of m is more informative than m. The latter informational comparison makes sense if belief functions are viewed as their equivalent credal set of probabilities, but the comparison of these intervals also makes sense in the framework of evidence theory. So we have a clash of intuitions since, while the consonant belief function induced by a likelihood function is the least committed among them, it is far from being the least informative among belief functions whose contour function is proportional to it.

This is one example of the divergence between the belief function and the imprecise probability settings on this statistical inference issue. In the imprecise probability setting, as shown in the paper of Masegosa and Moral [14], a likelihood function is viewed as the family of probabilities that generates it using Bayes conditioning. It brings very little (basically no) information on events of interest, if the prior probability is vacuous. What they call the *learning principle* consists in

assuming that the prior information is informative enough to ensure a non-vacuous update when new observations are acknowledged. Interestingly, sticking to the Bayes rule for the updating operation then leads these authors to assume a uniformly distributed prior, and imprecise probabilities only occur to reflect the inconsistency between the frequencies of further observations and this assumption. This drastic approach, which uses an arguably arbitrarily restrictive family of prior probabilities, for the sake of making the Bayesian updating non-trivial, totally contrasts with attempts to extract meaningful information from the likelihood functions only, in order to dispense with prior probabilities (such as the paper by Denoeux, but also the generalised Bayes theorem [19,17]; see also discussions in [13]). Resolving the dilemma between commonality and plausibility-based informational comparison in evidence theory may be of great interest in order to perform a fair comparative assessment of the use of belief functions and imprecise probability in statistics.

Another justification of the use of renormalised likelihood functions interpreted as possibility distributions was mentioned in [7], without using the minimal commitment principle. If we extend the likelihood $L(\theta) = cP(x|\theta)$ of elementary hypotheses to disjunctions thereof, it turns out that the corresponding set-function Λ should obey the laws of possibility measures [1,6] in the absence of probabilistic prior. Indeed the following properties look reasonable for such a set-function Λ:

- The properties of probability theory enforce $\forall T \subseteq \Theta$, $\Lambda(T) \leq \max_{\theta \in T} L(\theta)$;
- A set-function representing the likelihood of a disjunction of assumptions should be monotonic with respect to inclusion: If $\theta \in A \subseteq \Theta$, $\Lambda(A) \geq L(\theta)$;
- Keeping the same scale as probability functions, we assume $\Lambda(\Theta) = 1$.

Then it is clear that $L(\theta) = \frac{P(x|\theta)}{\max_{\tau \in \Theta} P(x|\tau)}$ and $\Lambda(A) = \max_{\theta \in A} L(\theta)$, i.e., the extended likelihood function is a possibility measure, and the coefficient c is then fixed. This is another way to recover Shafer's [16] proposal to define a consonant belief function induced by likelihood information. However it does not use the theory of belief functions, only possibility theory and imprecise probability.

4. On the consistency between models and interval data

The point made by Huellermeier [12], namely that the choice of a model that explains the data helps in disambiguating it, if imprecise, is quite different, and refers to the issue of determining the level of agreement between a class of models and a set of imprecise data. It all depends on how one considers the role of the imprecision of the data and how it should impact the result of the learning procedure.

A first possible concern is that of propagating the imprecision of the data over to result of the learning procedure via sensitivity analysis (an option discussed in [3]). Note that this method is similar to some estimation process, where, based on a given principle (such as maximum likelihood) an estimate is computed from the set of random data characterised by their probability distributions (instead of intervals); the estimate is then itself considered as a random variable whose distribution is obtained by propagating the data distributions through the mathematical model of the estimate. In the interval regression example, the regression line is considered as the optimal result obtained as a function of the data, and a sensitivity analysis is performed on this optimal result.

This view is questioned by Huellermeier [12] on the ground that among the various optimal models obtained by precise instances of the data, some of them look clearly more plausible than other ones despite the low quality data set. This consideration pertains to a concern that differs from sensitivity analysis. The thrust of the paper is to extend the loss minimisation approach to the determination of a precise model in order to handle imprecise data taking the form of intervals. In the Boolean case, the author considers the sum of minimal losses incurred when not hitting the intervals. This approach is extended to fuzzy intervals, integrating over alpha-cuts. Under this loss function, some pairs (precise data set, model) appear more likely than other ones, despite imprecision, which implies that a data disambiguation takes place (selecting the precise data values that achieve the loss-function minimisation).

Referring to Fig. 1 in [12], it seems that the reason why the regression line and the blue data set on the right-hand side picture of looks implausible is due to the particular choice of instances in each interval that contains a strong outlier influencing the positioning of the regression line. The existence of this outlier is due to a very large imprecision of this particular piece of data.

My view of the method advocated by Huellermeier is as follows: rather than visualising the impact of imprecision on the optimal regression line, the aim is to find the best *precise* model and data set instance in the sense of optimising its closeness with the interval data (as measured by the loss function). In the regression example, clearly, it seems that any affine function that would hit all intervals (if any) would qualify as an optimal result of Huellermeier's procedure (the loss function would be 0). The approach then leads to a model that is logically consistent with the interval data. This result would correspond to overfitting (hence dubious) if the data were precise. However, with imprecise data this is not the case, since on the one hand, this kind of perfect fit only indicates the possibility of good fit with the actual underlying dataset. On the other hand, the more imprecise the data the more frequent this situation is to be encountered (and the optimal models may not be unique any longer), while in the case of numerous precise data it is practically impossible.

More formally, we can describe this extreme situation as follows (referring to Fig. 1 as an example). Let $\mathbf{M}$ be the set of affine functions f standing for the assumed class of models, and $\mathcal{D} = \{(x_i, Y_i) : i = 1, \ldots, n\}$ be the set of interval data. $\mathbf{M}$ is

said to be consistent with $\mathcal{D}$ if $\exists f \in \mathbf{M}, \forall i = 1, \ldots, n, f(x_i) \in Y_i$. Let $\mathbf{M}_{\mathcal{D}}$ be the subclass of models consistent with $\mathcal{D}$. In this case, the disambiguation is simply due to the conjunctive fusion of $\mathbf{M}$ and $\mathcal{D}$, in the sense that each piece of data (x_i, Y_i) is finally made more precise and becomes (x_i, Y_i^*) where $Y_i^* = \{f(x_i) : f \in \mathbf{M}_{\mathcal{D}}\} \subseteq Y_i$. Note that the disambiguation is partial here, and the methodology comes down to a conjunctive fusion procedure: the use of a logical conjunction explains why a certain disambiguation takes place. Note that in that special situation, this approach can be viewed as a kind of interpolation method between samples of an interval-valued function.

A drastic understanding of this approach would reject the class of models M as incompatible with the data set $\mathcal{D}$ if the set $\mathbf{M}_{\mathcal{D}}$ is empty. The method of Huellermeier takes a milder, more realistic view. In the case of such incompatibility, the method computes the closest model in the sense prescribed by the loss function (just as in the least square procedure). As the resulting model f^* is unique, the data disambiguation would be complete.

If we assume that values close to the midpoint of the intervals are more likely than values close to their extremities, one may consider using e.g. triangular fuzzy intervals $\tilde{Y}_i$ instead of intervals Y_i. The membership grade $\mu_{\tilde{Y}_i}(y_i)$ then reflects the plausibility of the value y_i associated to x_i (it is a reward function rather than a loss function); and one could pose the regression problem as a fuzzy constraint satisfaction problem, namely: find $f \in \mathbf{M}$ that maximises $\min_{i=1}^{n} \mu_{\tilde{Y}_i}(f(x_i))$. Note that in [12] everything works as if every piece of interval data (x_i, Y_i) were enlarged to encompass the points $f^*(x_i)$ that do not lie in Y_i. It comes down to using fuzzy intervals whose cores are the interval data and to use $1 - \mu_{\tilde{Y}_i}(y_i)$ as local loss functions. However, in [12], the local loss functions are summed, contrary to the fuzzy constraint view.

This approach proposed in [12] also sounds relevant in data reconciliation methods, where data estimates must be modified so as to fit a prescribed model [15]. In that area, the least square approach is prominent and imprecise data are (often questionably) interpreted by means of a mean value and a standard deviation of a Gaussian distribution. See [5], where we address this problem by means of the fuzzy constraint-based method outlined above, in a material flow analysis problem, and for a discussion of the probabilistic approach to reconciliation problems. The reconciliation problem can be viewed as the dual of the learning problem: indeed, the thrust of the reconciliation problem is to learn the data based on the knowledge of the model, while the machine learning does the converse. The thought-provoking point made by Huellermeier is that the two problems can be simultaneously solved.

References

[1] G. Coletti, R. Scozzafava, Coherent conditional probability as a measure of uncertainty of the relevant conditioning events, in: Proc. of ECSQARU03, in: LNAI, vol. 2711, Springer-Verlag, 2003, pp. 407–418.

[2] A. Colubi, G. González-Rodríguez, M.Á. Gil, W. Trutschnig, Nonparametric criteria for supervised classification of fuzzy data, Int. J. Approx. Reason. 52 (9) (December 2011) 1272–1282.

[3] I. Couso, D. Dubois, Statistical reasoning with set-valued information: ontic vs. epistemic views, Int. J. Approx. Reason. 55 (7) (2014) 1502–1518, http://dx.doi.org/10.1016/j.ijar.2013.07.002.

[4] T. Denoeux, Likelihood-based belief function: justification and some extensions to low-quality data, Int. J. Approx. Reason. 55 (7) (2014) 1535–1547, http://dx.doi.org/10.1016/j.ijar.2013.06.007.

[5] D. Dubois, H. Fargier, D. Guyonnet, Data reconciliation under fuzzy constraints in material flow analysis (regular paper), in: Javier Montero, Gabriella Pasi, Davide Ciucci (Eds.), European Society for Fuzzy Logic and Technology Conference, EUSFLAT 2013, Milan, 11/09/2013–13/09/2013, Atlantis Press, 2013, pp. 25–32.

[6] D. Dubois, Possibility theory and statistical reasoning, Comput. Stat. Data Anal. 51 (1) (2006) 47–69.

[7] D. Dubois, T. Denoeux, Statistical inference with belief functions and possibility measures: a discussion of basic assumptions, in: International Conference on Soft Methods in Probability and Statistics, SMPS 2010, Oviedo (Spain), 28/09/2010–01/10/2010, in: Adv. Intell. Soft Comput., vol. 77, Springer, 2010, pp. 217–225.

[8] D. Dubois, H. Prade, A set-theoretic view of belief functions – logical operations and approximation by fuzzy sets, Int. J. Gen. Syst. 12 (3) (1986) 193–226.

[9] D. Dubois, H. Prade, Consonant approximations of belief functions, Int. J. Approx. Reason. 4 (5/6) (1990) 419–449 (Special Issue: Belief Functions and Belief Maintenance in Artificial Intelligence).

[10] D. Dubois, H. Prade, P. Smets, New semantics for quantitative possibility theory, in: Proc. of the 6th European Conference, ESQARU 2001, Toulouse, France, Springer-Verlag, 2001, pp. 410–421.

[11] G. González-Rodríguez, A. Colubi, M.A. Gil, Fuzzy data treated as functional data. A one-way ANOVA test approach, Comput. Stat. Data Anal. 56 (4) (2012) 943–955.

[12] Eyke Huellermeier, Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization, Int. J. Approx. Reason. 55 (7) (2014) 1519–1534, http://dx.doi.org/10.1016/j.ijar.2013.09.003.

[13] D. Dubois, A. Gilio, G. Kern-Isberner, Probabilistic abduction without priors, Int. J. Approx. Reason. 47 (3) (2008) 333–351.

[14] A.R. Masegosa, S. Moral, Imprecise probability models for learning multinomial distributions from data. Applications to learning credal networks, Int. J. Approx. Reason. 55 (7) (2014) 1548–1569, http://dx.doi.org/10.1016/j.ijar.2013.09.019.

[15] S. Narasimhan, C. Jordache, Data Reconciliation and Gross Error Detection: An Intelligent Use of Process Data, Gulf Publishing Company, Houston, 2000.

[16] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976.

[17] G. Shafer, Belief functions and parametric models, J. R. Stat. Soc. B 44 (1982) 322–352.

[18] P. Smets, Information content of an evidence, Int. J. Man-Mach. Stud. 19 (1) (1983) 33–43.

[19] P. Smets, Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem, Int. J. Approx. Reason. 9 (1) (1993) 1–35.

[20] P. Smets, The transferable belief model for quantified belief representation, in: P. Smets (Ed.), Handbook on Defeasible Reasoning and Uncertainty Management Systems – Volume 1: Quantified Representation of Uncertainty and Imprecision, Kluwer Academic Publ., Dordrecht, The Netherlands, 1998, pp. 267–301.

[21] SMIRE Research Group at the University of Oviedo, A distance-based statistical analysis of fuzzy number-valued data, Int. J. Approx. Reason. 55 (7) (2014) 1487–1501, http://dx.doi.org/10.1016/j.ijar.2013.09.020.